



Journal of Health Economics and Outcomes Research

Systematic Review



Faster but Not Truer, Part 1: A Systematic Scoping Review of Artificial Intelligence in Health Economics and Outcomes Research

Onur Baser^{1,2*}

¹Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA

²Graduate School of Public Health and Health Policy, City University of New York, New York, USA

ARTICLE INFORMATION

This is the first of two articles. Part 1 measures the literature and validates the measurement. Part 2 assembles the regulatory, health technology assessment, and industry record, examines the economics of automation as it applies to this field, and proposes the TRACE-HEOR reporting standard.

Accepted September 16, 2026

Keywords: artificial intelligence; machine learning; real-world evidence; health economics outcomes research; health technology assessment; causal inference

*Corresponding author:

Email address: onur.baser@sph.cuny.edu

➤ [Supplementary Material](#)

ABSTRACT

Background: Artificial intelligence (AI) is being introduced into every stage of the health economics and outcomes research (HEOR) production line, from literature screening to causal estimation. The field lacks a defensible count of what has been done, and bibliometric claims about AI adoption circulate without measured error rates.

Methods: Following PRISMA-ScR, we searched PubMed for AI terms co-occurring with HEOR terms for publications from 2010 through 2025, retrieved 2642 records, and included 2476 after screening. Records were classified into 8 nonexclusive application domains and 1 of 6 method categories using a published term dictionary. To quantify classification error, an independent large language model classifier, blinded to the dictionary labels and given only title and abstract, reclassified a random sample of 300 included records. We report per label counts, agreement, Cohen kappa, prevalence-adjusted bias-adjusted kappa, precision, recall, F1, and a precision-adjusted corpus size with exact binomial intervals. We separately assembled the primary performance evidence for each HEOR workflow, requiring a quantitative metric from a primary study.

Results: Included publications rose from 29 in 2010 to 852 in 2025, a factor of 29.4, with the inflection at 2018 to 2019 rather than at the release of ChatGPT. AI reached 6.0% of all indexed HEOR output in 2025 while HEOR fell to 0.86% of all indexed medical AI output. Economic evaluation was the most frequently assigned domain (1559/2476, 63.0%), followed by real-world data and causal inference (496, 20.0%) and governance, reporting, and validity (494, 20.0%). Generative AI appeared in no record before 2021 and in 103 in 2025. Only 154 of 300 sampled records (51.3%; 95% confidence interval [CI], 45.7-56.9) were judged genuinely HEOR relevant, implying a precision-adjusted corpus of 1271 studies (95% CI 1131-1410). Agreement was substantial for real-world data and causal inference (kappa, 0.729) but negligible for economic evaluation (kappa, 0.144), where dictionary precision was 29.6%. Of 2476 records, 38 (1.5%) were randomized trials and 17 (0.7%) validation studies. The replication package is publicly available.

Conclusions: The literature is growing faster than its own measurement infrastructure. Measured gains concentrate in tasks with observable ground truth, and predictive performance repeatedly fails to predict evidentiary performance.

INTRODUCTION

Two facts about artificial intelligence (AI) in health economics and outcomes research (HEOR) sit uneasily beside each other. This series is about the distance between them.

The first is that the field has been reorganized around it. Since 2016, the US Food and Drug Administration (FDA) has received more than 500 drug and biological product submissions containing AI

components.¹⁻³ The agency issued its first draft guidance on AI for regulatory decision making in January 2025,⁴ joined the European Medicines Agency in publishing 10 guiding principles for good AI practice in drug development in January 2026,^{5,6} and has now authorized 1614 AI-enabled medical devices.⁷ The International Society for Pharmacoeconomics and Outcomes Research moved AI from the third-ranked HEOR trend for 2024 to 2025 to the first-ranked trend for 2026 to



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

2027,⁸ and published a taxonomy, a reporting guideline, and a policy report for generative AI in HEOR within a single year.⁹⁻¹¹

The second is that almost none of it has reached the documents on which reimbursement decisions are made. In a systematic audit of 1391 health technology assessment (HTA) documents submitted to 3 agencies between 2023 and 2025, exactly 3 disclosed the use of AI or machine learning, an adoption rate of 0.2%.¹² When editors of *The Lancet* checked every submission received during February 2024, 0.6% declared use of ChatGPT.¹³

That gap is a measurement problem before it is anything else, and the measurement problem is the subject of Part 1. We do not have a defensible count of what has actually been done. Prior reviews have been deliberately partial. Lee et al screened 148 records for wearable applications and included 32,¹⁴ then sampled 161 of 805 records for nonwearable applications and included 92.¹⁵ A 2026 scoping review of the economic aspects of machine learning in health care screened 6332 references and included 18,¹⁶ and a protocol for a review of machine learning in health economic evaluation screened 4141 records and found 30 eligible reports.¹⁷ Each answers a narrower question than the one practitioners now ask, which is how large the field is, what it is actually doing, and how fast its composition is changing.

The reason the answer matters beyond bibliometrics is that the standards now being written for AI in evidence generation are being written against an evidence base whose reliability nobody has quantified. All the current frameworks are risk based, and all of them presume the submitter can characterize the risk. Yet the prediction model literature they inherit is dominated by high risk of bias: 545 of 606 COVID-19 prediction models were rated at high risk of bias,¹⁸ as were 148 of 171 studies of machine learning–based prediction models.¹⁹

This article therefore does three things. First, it reports an original systematic scoping review of the indexed literature following PRISMA-ScR, covering 2010 through 2025, with a full and reproducible search rather than a sample. Second, and unusually for a scoping review, it validates its own classification against an independent blinded classifier and reports the resulting error rates, because a bibliometric claim without a measured error rate is an assertion. Third, it sets the measured performance evidence for each HEOR workflow beside that map, so that the size and the quality of a literature can be read together. Part 2 takes up the regulatory, health technology assessment, and industry record, the economics of automation as it applies to this field, and a proposed reporting standard. The intended readers are the people who will have to defend this evidence: HEOR scientists and biostatisticians in industry and academia, HTA methodologists writing the next generation of position statements, and reviewers who will receive dossiers built with these tools whether the dossiers say so.

METHODS

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR). The protocol, all search strings, the classification dictionary, and the analysis code are archived in a public replication package.²⁰ No patient level data were used, and no ethics approval was required.

Search Strategy and Eligibility

We searched PubMed on 9 September 2026 using a structured Boolean query requiring co-occurrence of at least 1 AI term and at least 1 HEOR term in the title, abstract, or MeSH terms, restricted to publication years 2010 through 2025. AI terms covered *artificial intelligence*, *machine learning*, *deep learning*, *neural networks*, *natural language processing*, *large language models*, *generative AI*, and *named model families*.

HEOR terms covered *cost effectiveness*, *cost utility*, *economic evaluation*, *decision analytic modeling*, *budget impact*, *health technology assessment*, *real world evidence*, *comparative effectiveness*, *patient reported outcomes*, *utilities*, *quality adjusted life years*, and *health care resource utilization*. The complete query strings are reproduced in **Supplementary Table S1**.

Records were eligible if they were journal articles, reviews, or conference proceedings published in English between 2010 and 2025 with a retrievable abstract. We excluded editorials, letters, comments, preprints, retracted publications, and records dated outside the window, and removed duplicates on PubMed identifier and normalized title.

To place the corpus in context, we retrieved annual denominator series from PubMed using the same interface, reported in **Supplementary Table S5**. One discrepancy should be noted. The range form of the query returned 2656 records where the sum of the 16 per-year queries was 2929, a difference of 273, most likely from MedlineDate records counted in 2 calendar years. We used the per-year series for trends and the range query for the PRISMA flow.

Classification

Each included record was classified by a deterministic term dictionary applied to the concatenated title, abstract, and MeSH terms. The dictionary assigns 8 nonexclusive application domains: economic evaluation; decision modeling and simulation; evidence synthesis; real-world data and causal inference; outcome and cost prediction; patient-reported outcomes and preferences; HTA, payer, and market access; and governance, reporting, and validity. It separately assigns a single most specific method category from six: generative AI and large language models; natural language processing before the generative era; deep learning; classical machine learning; artificial intelligence–unspecified; and no AI method term identified. Domain labels are nonexclusive because a single paper frequently reports, for example, a natural language processing pipeline used to construct a cohort for a cost outcome. The full dictionary is provided in **Supplementary Table S2**.

Validation of the Classifier

A dictionary is fast, transparent, and reproducible, and it is also wrong in ways that are invisible unless measured. We drew a simple random sample of 300 of the 2476 included records and had an independent large language model classifier, blinded to the dictionary output and given only title and abstract, answer 2 questions: whether the record was genuinely relevant to HEOR as a practitioner would define it, and which of the same 8 domains and 6 method categories applied. Prompts and raw outputs were archived.

We reported, for every label, the number assigned by each classifier, observed agreement, Cohen kappa, prevalence-adjusted bias-adjusted kappa (PABAK), and the precision, recall, and F1 of the dictionary treating the blinded classifier as the reference. PABAK was reported alongside kappa because kappa is unstable at extreme prevalence and several labels are rare. We also reported search precision, defined as the proportion of sampled records judged genuinely HEOR relevant, with an exact binomial 95% confidence interval (CI), and a precision-adjusted corpus size obtained by applying that proportion and its interval to the full included count. We treated the blinded classifier as a reference standard, not as truth, and returned to that choice in the Limitations.

Narrative Evidence Review

Alongside the bibliometric review we assembled a source grounded evidence base on the measured performance of AI in each HEOR workflow, requiring a primary study reporting a quantitative metric. Where a figure could not be confirmed from a retrieved page, we recorded it as not available rather than estimating it. Gray literature without a

digital object identifier was cited with the issuing organization, date, and uniform resource locator.

Statistical Analysis

Counts were reported with percentages of the relevant denominator and CIs for proportions are exact binomial (Clopper-Pearson). Growth was expressed as a ratio of endpoint counts. No hypothesis tests were reported for the bibliometric series because the series are complete counts rather than samples. Analyses used Python 3.11, with versions pinned in the replication package.

RESULTS

Size and Shape of the Literature

The search identified 2642 records. After removal of 14 duplicates, 2628 were screened and 152 excluded: 70 without a retrievable abstract, 38 not in English, 19 editorials, 10 preprints, 7 letters, 4 retracted, and 4 with a publication year outside the window. This left 2476 included records. Figure 1 presents the PRISMA-ScR flow, itemized in Supplementary Table S3, and Table 1 summarizes the corpus.

Annual output rose from 29 records in 2010 to 852 in 2025, a factor of 29.4, with the inflection at 2018 to 2019 rather than at the release of ChatGPT. The series was 29, 28, 47, 31, 34, 24, 23, 36, 48, 91, 130, 163, 232, 270, 438, and 852, the largest proportional increase being the rise of 89.6% from 2018 to 2019. Figure 2 shows the series against its denominators and Supplementary Table S4 gives counts per year by domain.

Two ratios matter more than the raw count. AI containing publications reached 6.0% of all indexed HEOR output in 2025, having been 0.8% in 2010, while HEOR fell to 0.86% of all indexed medical

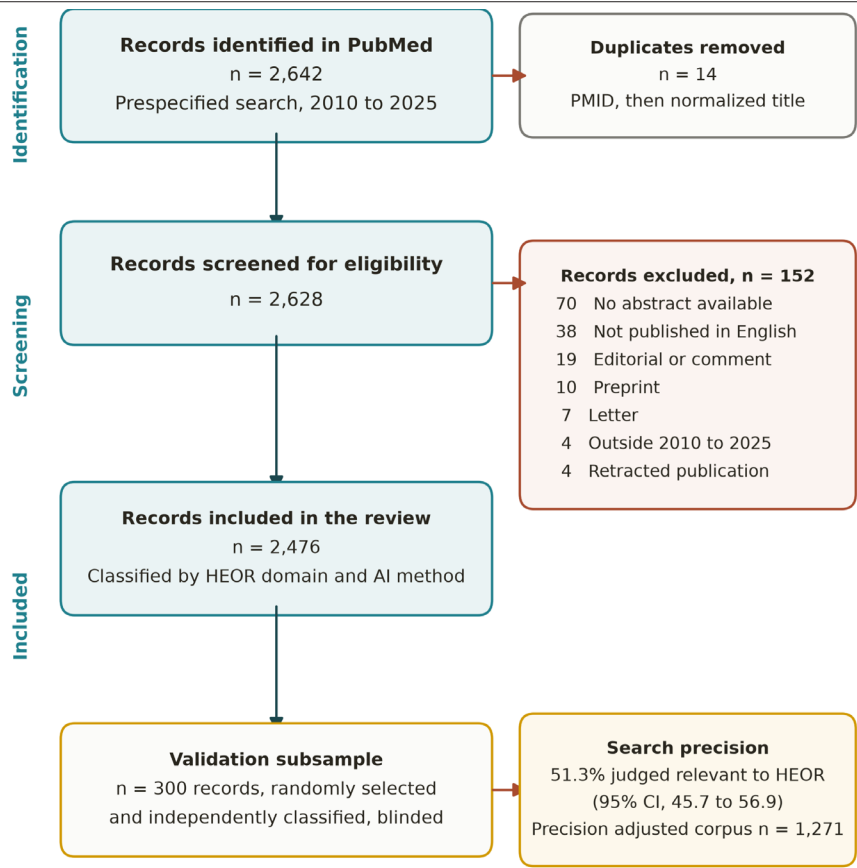
AI output, since PubMed indexed 110494 AI in medicine records against 15641 HEOR records that year. That asymmetry means HEOR will be a consumer of externally developed tools far more often than a producer of them, inheriting validation practices set elsewhere.

Domain composition is reported in Table 1 and mapped in Figure 3. Economic evaluation was the most frequently assigned domain, appearing in 1559 of 2476 records (63.0%), followed by real-world data and causal inference (496, 20.0%), governance, reporting, and validity (494, 20.0%), outcome and cost prediction (393, 15.9%), evidence synthesis (283, 11.4%), HTA, payer, and market access (209, 8.4%), decision modeling and simulation (115, 4.6%), and patient-reported outcomes and preferences (76, 3.1%). No domain term was assigned in 278 records (11.2%). Method composition was artificial intelligence–unspecified in 682 records (27.5%), classical machine learning in 615 (24.8%), deep learning in 445 (18.0%), no identifiable AI method term in 445 (18.0%), generative AI or large language models in 159 (6.4%), and pre-generative natural language processing in 130 (5.3%).

The generative series is empty before 2021, then 1, 4, 13, 38, and 103 records for 2021 through 2025. Growth is steep and the level is low: even in 2025 generative AI appeared in 12.1% of output. The fastest growing domain label in absolute terms was governance, reporting, and validity, which rose from 82 records in 2024 to 237 in 2025. The field is now writing more about how to govern these tools than it is publishing indexed applications of them.

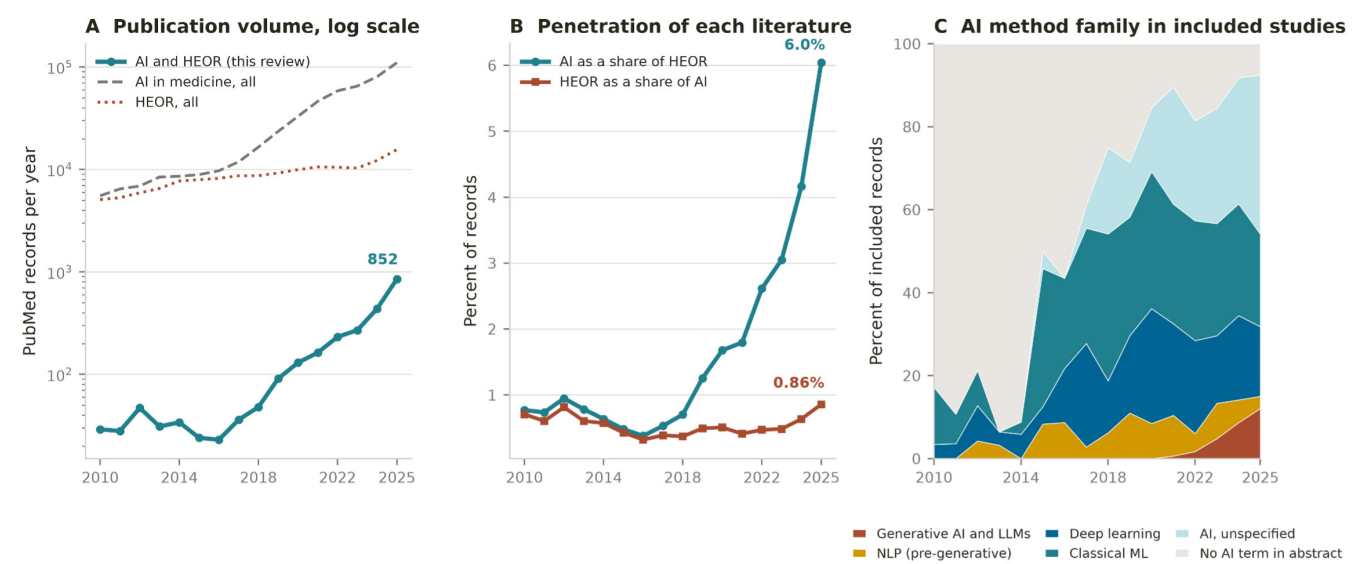
Venues were highly dispersed across more than 900 distinct journals, led by *Scientific Reports* (79 records, 3.2%), *Sensors* (60, 2.4%), *Cureus* (43, 1.7%), *Journal of Medical Internet Research* (41, 1.7%), and *Value in Health* (32, 1.3%). That the field’s own specialty journal accounts for 1.3% of the corpus is the important observation: the

Figure 1. PRISMA-ScR Flow of Study Identification, Screening, and Inclusion



Records identified in PubMed for publication years 2010 through 2025, with duplicate removal, itemized exclusions, and the 2476 included records. The precision-adjusted corpus estimate derived from blinded independent classification of 300 randomly sampled included records is shown for comparison.

Figure 2. Growth and Composition of the AI and HEOR Literature, 2010-2025, Against Its Denominators



A: Publication volume on a logarithmic scale for the included corpus, for all indexed medical AI publications, and for all indexed HEOR publications.

B: Penetration of each literature into the other (ie, AI-containing HEOR publications as a percentage of all indexed HEOR publications, which reached 6.0% in 2025, and HEOR publications as a percentage of all indexed medical AI publications, which reached 0.86%).

C: The method family composition of the included corpus as a percentage of records per year, with generative AI and large language models, pregenerative natural language processing, deep learning, classical machine learning, artificial intelligence–unspecified, and records with no AI method term in the abstract.

Table 1. Characteristics of the Included Corpus (N = 2476)

Characteristic	Included Records (N)	% of Included	Classification Reliability
Publication year			—
2010	29	1.17	—
2011	28	1.13	—
2012	47	1.90	—
2013	31	1.25	—
2014	34	1.37	—
2015	24	0.97	—
2016	23	0.93	—
2017	36	1.45	—
2018	48	1.94	—
2019	91	3.68	—
2020	130	5.25	—
2021	163	6.58	—
2022	232	9.37	—
2023	270	10.90	—
2024	438	17.69	—
2025	852	34.41	—
Application domain (nonexclusive)			Cohen kappa vs blinded classifier
Economic evaluation	1559	62.96	0.144
Real-world data and causal inference	496	20.03	0.729
Governance, reporting, and validity	494	19.95	0.546
Outcome and cost prediction	393	15.87	0.453
Evidence synthesis	283	11.43	0.666
HTA, payer, and market access	209	8.44	0.576
Decision modeling and simulation	115	4.64	0.579
Patient-reported outcomes and preferences	76	3.07	0.306
No domain term identified	278	11.23	—

Table 1. Characteristics of the Included Corpus (N = 2476), *cont'd*

Characteristic	Included Records (N)	% of Included	Classification Reliability
Method category (most specific, exclusive)			
Artificial intelligence–unspecified	682	27.54	0.667 (6-category multiclass)
Classical machine learning	615	24.84	–
Deep learning	445	17.97	–
No AI method term identified	445	17.97	–
Generative AI and large language models	159	6.42	–
Natural language processing (pre-generative)	130	5.25	–
Leading journals			
<i>Scientific Reports</i>	79	3.19	–
<i>Sensors (Basel, Switzerland)</i>	60	2.42	–
<i>Cureus</i>	43	1.74	–
<i>Journal of Medical Internet Research</i>	41	1.66	–
<i>BMJ Open</i>	36	1.45	–
<i>PloS One</i>	35	1.41	–
<i>Value in Health</i>	32	1.29	–
<i>Journal of Clinical Medicine</i>	27	1.09	–
<i>NPJ Digital Medicine</i>	26	1.05	–
<i>Frontiers in Public Health</i>	20	0.81	–
Indexed publication type			
Journal article	2471	99.80	–
Review	674	27.22	–
Research support, non-US government	492	19.87	–
Systematic review	121	4.89	–
Research support, NIH, extramural	105	4.24	–
Comparative study	91	3.68	–
Multicenter study	54	2.18	–
Scoping review	42	1.70	–
Randomized controlled trial	38	1.53	–
Research support, US government, non-PHS	37	1.49	–
Clinical trial protocol	31	1.25	–
Meta-analysis	27	1.09	–

literature HEOR practitioners will be asked to appraise is overwhelmingly published outside HEOR venues and therefore reviewed by people who do not routinely apply HEOR reporting standards. **Supplementary Table S6** lists the 40 most frequent journals.

Study designs were correspondingly weak on the dimensions HTA reviewers care about. Of 2476 records, 674 (27.2%) were indexed as reviews and 121 (4.9%) as systematic reviews, against 38 (1.5%) randomized controlled trials and 17 (0.7%) validation studies. **Supplementary Table S7** gives the full distribution.

What Validation Reveals

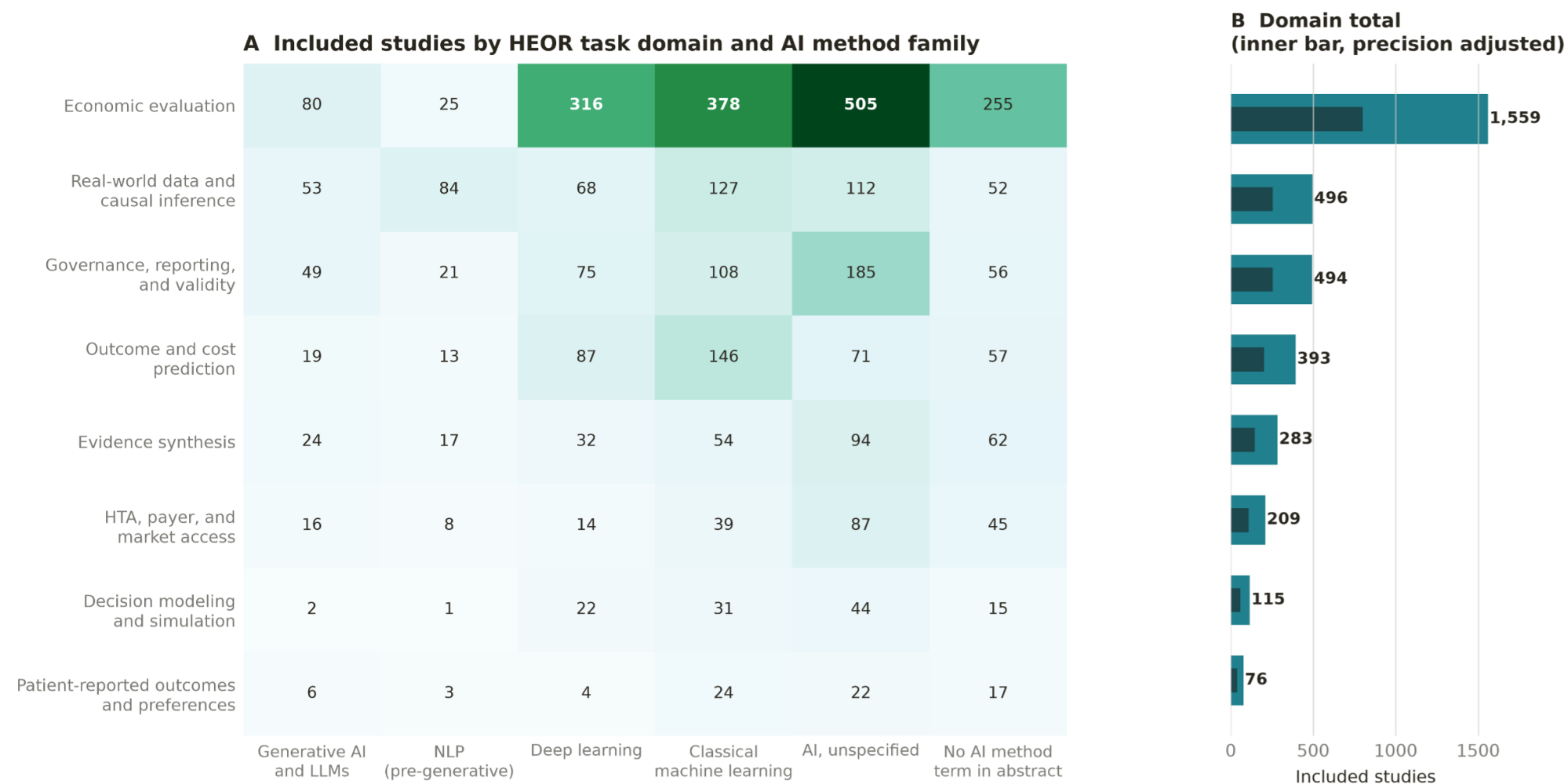
The blinded independent classification of 300 randomly sampled records produced the most consequential result in this review, and it is not a flattering one.

Only 154 of 300 sampled records (51.3%; 95% CI, 45.7–56.9) were judged genuinely relevant to health economics and outcomes research. Applying that proportion to the corpus gives a precision-adjusted estimate of 1271 genuinely HEOR relevant studies (95% CI, 1131–1410) rather than 2476. Precision was 52.4% (95% CI, 46.5–58.3) among the 271 sampled records to which the dictionary had assigned at least 1 domain and 41.4% (95% CI, 25.5–59.3) among the

29 to which it had assigned none, so the dictionary carries signal, but not nearly enough of it. **Table 2** and **Figure 4** report the validation in full, with the stratified precision in **Supplementary Table S9**.

Per-label agreement was heterogeneous in a way that maps almost exactly onto how well specified each concept is. Agreement was substantial for real-world data and causal inference (Cohen kappa, 0.729; observed agreement, 90.3%; F1, 79.1%) and moderate to substantial for evidence synthesis (kappa, 0.666; F1, 70.9%), decision modeling and simulation (kappa, 0.579), HTA, payer, and market access (kappa, 0.576), and governance, reporting, and validity (kappa, 0.546). It was fair for outcome and cost prediction (kappa, 0.453) and poor for patient-reported outcomes and preferences (kappa, 0.306, F1, 33.3%). Method classification agreed at kappa 0.667 across 6 categories and the generative AI flag at kappa 0.674. Full statistics are in **Supplementary Table S8**.

The exception should change how bibliometric claims about this field are read. For economic evaluation, the largest and fastest growing label in the corpus, Cohen kappa was 0.144 with observed agreement of 51.7% and PABAK of 0.033. The dictionary assigned this label to 186 of 300 sampled records; the blinded classifier assigned it to 69. Dictionary precision was 29.6% with recall 79.7%. In plain terms, the

Figure 3. Evidence Map of AI Applications Across HEOR Tasks

A: Cross-tabulation of the 8 application domains against the 8 method families; cell values are counts of included records and domain labels are nonexclusive, so rows do not sum to the corpus.

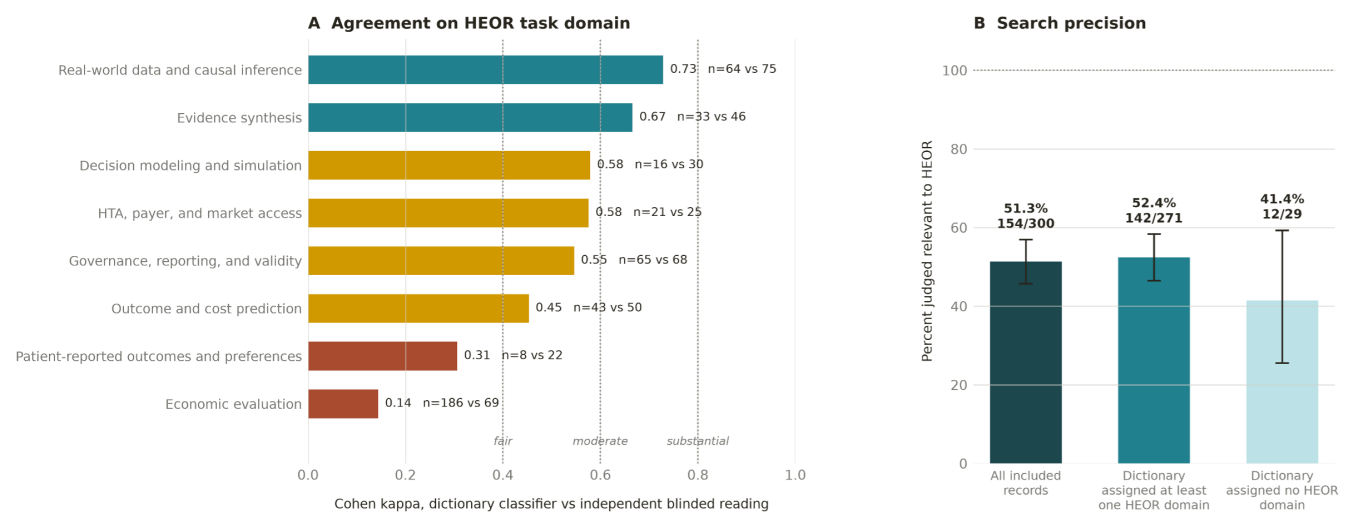
B: The domain total for each row, with the precision-adjusted total shown as an inner bar so that the measured and the nominal size of each literature can be read together.

Table 2. Validation of the Classification Against Blinded Independent Classification of 300 Randomly Sampled Records

Label	Assigned by Dictionary (n)	Assigned by Blinded Classifier (n)	Observed Agreement (%)	Cohen Kappa	PABAK	Precision (%)	Recall (%)	F1 (%)
Economic evaluation	186	69	51.7	0.144	0.033	29.6	79.7	43.1
Decision modeling and simulation	16	30	94.0	0.579	0.880	87.5	46.7	60.9
Evidence synthesis	33	46	92.3	0.666	0.847	84.8	60.9	70.9
Real-world data and causal inference	64	75	90.3	0.729	0.807	85.9	73.3	79.1
Outcome and cost prediction	43	50	85.7	0.453	0.713	58.1	50.0	53.8
Patient-reported outcomes and preferences	8	22	93.3	0.306	0.867	62.5	22.7	33.3
HTA, payer, and market access	21	25	94.0	0.576	0.880	66.7	56.0	60.9
Governance, reporting, and validity	65	68	84.3	0.546	0.687	66.2	63.2	64.7
Method category (6-category multiclass)	300	300	73.7	0.667				
Generative AI or LLM flag			97.0	0.674	0.940			

Search Precision	Sampled (n)	Judged HEOR-Relevant (n)	Precision (%)	95% CI Lower	95% CI Upper
All included records	300	154	51.3	45.7	56.9
Dictionary assigned ≥1 HEOR domain	271	142	52.4	46.5	58.3
Dictionary assigned no HEOR domain	29	12	41.4	25.5	59.3
Precision-adjusted corpus size	2476 included	1271		1131	1410

Per-label counts, observed agreement, Cohen kappa, PABAK, precision, recall, and F1, with search precision and the precision-adjusted corpus estimate.

Figure 4. Validation of the Term Dictionary Against Blinded Independent Classification of 300 Randomly Sampled Included Records

A: Cohen kappa for agreement on each HEOR task domain, annotated with the number of records the dictionary and the blinded classifier each assigned to that domain, and shaded by the conventional interpretive bands for fair, moderate, and substantial agreement.

B: Search precision overall and stratified by whether the dictionary assigned at least 1 domain label or none, with exact binomial 95% CIs. Full agreement statistics, including prevalence-adjusted bias-adjusted kappa, precision, recall, and F1, are given in **Table 2** and **Supplementary Table S8**.

dictionary treats any paper that mentions cost, resource use, or efficiency as an economic evaluation, and roughly 7 in 10 of those are not economic evaluations in the sense a reimbursement reviewer would use.

Two implications follow. First, our own headline count of 1559 economic evaluation records is an upper bound with an implied true value near 574 by the precision ratio, and any bibliometric claim about the growth of AI in health economic evaluation that does not report a measured classification error rate should be treated the same way. Second, the ranking of kappa values is a diagnostic of conceptual clarity in the field itself: real-world data and causal inference has a shared vocabulary and scores well, patient-reported outcomes and preferences does not and scores worst.

Evidence Synthesis: The Best Measured Workflow

Evidence synthesis is where the performance evidence is strongest and where it is most consistently misread. **Supplementary Table S10** lists the screening studies. The best result in the literature is Matsui et al, who applied GPT-4 to 1381 and 3146 record sets and reported adjusted sensitivity of 0.962 and specificity of 0.996 at about 110 records per minute and a cost of US\$59.²¹ It has not been reproduced at scale independently. Tran et al, on 22665 citations, reported sensitivity between 81.1% and 96.5% but specificity between 25.8% and 80.4%, generating 10,279 false positives in the worst configuration.²² Oami et al, on 16669 citations, reported sensitivity of 0.85 and specificity of 0.98 but a median specificity gap of 0.48 across reviews.²³ Khraisha et al found GPT-4 sensitivity of 0.42 with kappa 0.34 against human reviewer kappa of 0.77.^{24,25}

Read together, these say something actionable: sensitivity is acceptable, specificity is unstable across topics, and the variance between reviews exceeds the difference between models. That is the profile of a second screener with human adjudication, not of a sole screener. It is the conclusion of the joint position statement of Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence,²⁶ and why the ISPOR agentic framework reports screening F1 between 0.917 and 0.977 under a human in the loop design rather than an autonomous one.²⁷

Data extraction performs better than screening, and **Supplementary Table S11** gives the studies. Gartlehner et al extracted 160 items

with Claude 2 at 96.3% accuracy,²⁸ and Liu et al reported 94.8% accuracy at 88 seconds per randomized trial.²⁹ The most informative study ran 12814 extractions across 107 randomized trials and found Claude 3.5 at 96.2% accuracy against human extractors at 95.3% and human extraction assisted by the model at 97.9%, with extraction time falling from 86.9 minutes to 14.7.³⁰ The point is not that the model beat the humans, since the difference sits inside plausible noise. The point is that the combination beat both.

Downstream analytic steps automate as well. Reason et al had GPT-4 write 30 R scripts to replicate published cost effectiveness models; in the non-small cell lung cancer set 14 of 15 scripts were error free, and every error free script reproduced the published incremental cost effectiveness ratio to within 1%.^{31,32} A systematic review of generative AI in health economic evaluation screened 2322 records and included 25 studies, of which 92% were conference abstracts and 96% were company authored.³³ That composition is the finding.

The economic case rests on a smaller base than the enthusiasm suggests. Abogunrin et al found 17 of 25 studies reporting time reductions above 50%, against a per review cost above US \$141 000.^{34,35} The most honest comparison available is the ISPOR time and motion study of 234 publications: a hybrid human and AI workflow achieved 97.0% accuracy with 80.6% recall and 51.4% time savings, a fully automated workflow achieved 81.1% time savings with recall of only 48.4%, and unaided humans achieved 90.1% accuracy with 88.7% recall.²⁷ A workflow that misses half the eligible evidence is not 81% cheaper. It is a different study.

Real-World Data and Causal Inference

This is the domain our validation ranked most coherent, and the domain where the evidence most directly contradicts the intuition that better prediction yields better inference. **Supplementary Table S12** maps the methods.

The foundational result is Schneeweiss et al on high-dimensional propensity scores, where the estimated relative risk in a cohort of 49653 users moved from 1.09 crude to 0.94 with 15 investigator specified covariates and to 0.88 with 500 algorithmically selected covariates.³⁶⁻³⁸ Benasseur et al showed the failure mode plainly: high-dimensional propensity score weighting produced bias of 0.094 with 23.6%

CI coverage where a modified procedure achieved bias of 0.018 with 91.4% coverage.³⁹ Coverage of 23.6% for a nominal 95% interval is not a subtle problem.

The machinery is mature. Targeted maximum likelihood estimation,⁴⁰ super learning,⁴¹ causal forests,⁴² and double or debiased machine learning⁴³ are in routine use, and their translation into HEOR has been mapped through causal forests for heterogeneity in cost effectiveness analysis,^{44,45} integration of decision modeling with machine learning based causal estimation,⁴⁶ a scoping review yielding 74 studies of heterogeneous treatment effect estimation with real-world data,⁴⁷ and transportability methods.⁴⁸⁻⁵⁰

What the field lacks is calibration of expectations, and the clearest demonstration is a trial emulation. Wang et al emulated PARADIGM-HF in Veterans Affairs data, where the randomized hazard ratio was 0.81 (95% CI, 0.61-1.06). A logistic propensity score model with area under the curve of 0.79 yielded 0.93; a gradient-boosted model with area under the curve of 0.95 yielded 0.61.⁵¹ The model with the best discrimination produced the estimate furthest from the trial benchmark. Propensity score models are not prediction problems, and optimizing separation of treated from untreated can degrade the covariate balance that identification requires. It is, however, exactly the inference a reviewer would draw from an area under the curve reported without a balance diagnostic, and it is why FDA guidance on real-world data declines to endorse any specific AI technology,⁵² why the target trial framework has been recommended as the organizing discipline,⁵³ and why the PALISADE checklist was written.⁵⁴

Unstructured Data and AI-Derived Endpoints

Approximately 80% of clinical data is unstructured, and this is the strongest affirmative case for AI in HEOR.⁵⁵ The gains are not marginal. Social and behavioral determinants recorded in claims at prevalences between 0.03% and 8.3% were recorded at between 1.40% and 36.3% once notes were processed, with lack of social support identifiable only from text in 99.8% of cases.⁵⁶ In an oncology cohort, 44 of 48 patients (91.7%) with adverse social determinants documented in text had no corresponding Z code.⁵⁷

Performance on specific extraction tasks is good and uneven, as **Supplementary Table S13** shows. Xu et al, across 79 698 patients and 495 862 notes, an achieved area under the receiver operating characteristic curve of 0.95 for performance status and found that imputed poor performance status carried a hazard ratio of 11.9.⁵⁸ A Veterans Affairs implementation reported 98.5% accuracy⁵⁹ and programmed death ligand 1 extraction reached micro F1 of 0.925,⁶⁰ while one commercial oncology model reported precision of 98.8% for stage value but recall of only 79% for linking histology to neoplasm,⁶¹ and in another registry task T stage extraction reached F1 of 0.89 while M stage reached 0.24.⁶² The pattern is that the variable a cohort definition depends on is often the variable extracted worst.

Two results show why this matters for an estimate rather than for a metric. Varma et al derived real-world progression-free survival two ways in the same cohort and obtained 30 months from radiology reports against 23 months from clinical notes, with median real-world progression-free survival of 20 months (95% CI, 18-25) against 25 months (95% CI, 15-35) by manual curation in a validation subset.⁶³ A 7-month difference in a median attributable to which document type was read is a difference that changes a cost effectiveness result. Separately, adding natural language processing features raised propensity score model area under the curve from 0.66 to 0.77 and from 0.80 to 0.89 while the estimated hazard ratios barely moved, which is the PARADIGM-HF lesson in another guise.

Vendor validation frameworks have begun to respond. Flatiron Health published the VALID framework, requiring variable-level

benchmarking against expert human abstraction, internal consistency checks, replication, and subgroup stratified bias assessment,^{64,65} and its machine learning extracted real-world response correlated with human abstraction at $r = 0.86$ in a test subset of 4047.⁶⁶⁻⁶⁸ Verantos and coauthors from Amgen, Johnson and Johnson, and GSK reported recall improving from 49.9% to 93.2% and completeness from 46.1% (95% CI, 38.2-54.0) to 96.6% (95% CI, 85.8-107.4).⁶⁹ A reported CI with an upper bound of 107.4% for a proportion is itself a reporting problem, and all of these validation efforts were authored by parties with a commercial interest in the result.⁷⁰

The decisive gap is regulatory. FDA has accepted electronic health record-derived real-world data in a label expansion, approving palbociclib for men with advanced breast cancer on the basis of records from three databases.⁷¹ That establishes acceptance of an electronic health record derived source, not acceptance of an AI derived endpoint, and we were unable to document a single instance of any regulator or HTA body explicitly accepting or rejecting one as such. For a sponsor deciding whether to build a submission on real-world progression, that absence is the single most important fact in this series.

Decision Modeling, Simulation, and Synthetic Data

Decision modeling is the smallest domain (115 records, 4.6%) and contains the cleanest efficiency results, because its tasks are computational rather than inferential. Metamodeling reduced an expected value of partial perfect information calculation requiring 12.5 million model runs from more than 1000 days of brute force computation to about two days,⁷² and automated model coding replicated published incremental cost effectiveness ratios to within 1%.^{11,31,73-75}

Decision modeling is also the part of HEOR where uncertainty is already formalized, and therefore the part where AI is most easily disciplined. Otten et al cataloged 80 methods for assessing uncertainty in decision analytic models, nearly all of which presuppose that the uncertainty has been quantified,^{76,77} and that presupposition is what an opaque model breaks. Value of information analysis⁷⁸⁻⁸¹ and the CHEERS 2022 requirement that inputs and their sources be reported⁸² define what an AI-assisted model must still deliver.

Synthetic data is where the promise is largest and the reporting weakest, and **Supplementary Table S14** collects the studies. The utility evidence is genuine: D'Amico et al generated synthetic data for 7133 patients with clinical fidelity of 93% and genomic fidelity of 90% and constructed a synthetic control arm of 187 matched patients,^{83,84} and Wang et al built AI-generated digital twins from 6736 subjects in Alzheimer disease, in which 6236 of 6728 subjects with recorded race were White and 152 were Black,⁸⁵ a composition that determines what the resulting population can be used to claim.

Generator choice dominates the result. Akiya et al found that classification and regression tree-based generation produced synthetic data sets whose median progression-free survival fell inside the real 95% CI in 88.8% to 98.1% of replicates, against 6.5% to 37.8% for a conditional tabular generative adversarial network.⁸⁶ A generator that recovers the target quantity in 6.5% of replicates is not a lower quality version of one that recovers it in 98%; it is unusable for that purpose, and nothing in the summary statistics of the synthetic data would tell you which one you had. Reporting has not kept pace: of 73 studies, 70 (95%) evaluated utility while only 31 of 67 (46%) claiming a privacy purpose evaluated privacy, and only 8 of 73 (11%) used a differentially private generator.⁸⁷

DISCUSSION

Two findings organize this first article, and both are about measurement. The first is that the size of this literature cannot be read from

a keyword count, which overstates it by a factor of two. A corpus of 2476 records reduces to approximately 1271 genuinely HEOR-relevant studies (95% CI, 1131-1410) under blinded independent classification, and the label that grew fastest, economic evaluation, is the label our own classifier got most wrong (Cohen kappa, 0.144, precision 29.6%). A count without an error rate is a marketing figure, and if a dictionary built specifically for this task and published in full achieves precision of 29.6% on the single most important label, then keyword-based counts circulating in conference abstracts and market reports should be read as upper bounds of unknown looseness.

The second is a methodological finding that recurs in three independent places and should be stated as a rule. In the PARADIGM-HF emulation, a propensity model with area under the curve of 0.95 produced a hazard ratio of 0.61 against a trial value of 0.81, while a weaker model at 0.79 produced 0.93.⁵¹ In the natural language processing-augmented propensity work, discrimination rose substantially while the effect estimates barely moved. In drift monitoring across 4 deployed systems, calibration degraded before discrimination did.⁸⁸ The rule is that predictive performance is not evidentiary performance: HEOR quantities are calibrated, counterfactual, and decision relevant, and none of those properties is measured by a receiver operating characteristic curve. A dossier that reports model performance without reporting the sensitivity of the decision to that model has reported the wrong thing.

The two findings share a structure. Where a task has an observable answer, the gains are real, replicated, and often large: extraction accuracy against human abstraction at 96.2% against 95.3%,³⁰ screening recall against a human reference,²⁷ code generation against a published incremental cost effectiveness ratio,³¹ and computation against brute force.⁷² Where the answer is not observable, there is no reported instance of AI improving the accuracy of a cost effectiveness estimate, because accuracy of a cost effectiveness estimate is not measurable. The kappa ranking in our own validation reproduces the same division from the outside.

The consequence for practice is that the two layers should be governed differently. Clerical automation needs throughput and audit trails. Inferential use needs a decision-level sensitivity analysis, because nothing else in the toolkit will detect the failure. Part 2 takes up what regulators and HTA agencies have actually said about this division, what industry reports, and what a reporting standard built for it would require.

Limitations

We searched a single database. PubMed does not index much of the conference and gray literature where AI in HEOR is most active, and 92% of the studies in one review of generative AI in health economic evaluation were conference abstracts,³³ so our corpus understates operational activity.

Our classifier is a term dictionary, and its errors are large, measured, and reported rather than hidden. Precision against blinded independent classification was 51.3% overall and 29.6% for economic evaluation, so all domain counts should be read as upper bounds.

Our reference standard is itself a large language model rather than a human expert panel. This is a genuine circularity in an article about the reliability of these tools, and it cuts both ways: the blinded classifier may be systematically strict about what counts as HEOR, in which case our precision estimate is too low, or lenient on unfamiliar subfields, in which case some labels are understated. A human dual review of the same 300 records is the correct next step, and the sample and both label sets are archived so that anyone may perform it.²⁰

Our denominator series contains an unresolved discrepancy of 273 records between the range query and the sum of per-year queries,

and we restricted to English-language records, which biases the corpus toward North American and European practice.⁸⁹

CONCLUSIONS

The literature on AI in HEOR has grown by a factor of 29 since 2010 and reached 852 indexed records in 2025. Under blinded independent classification, approximately half of it is about something else, and the fastest growing label is the one that can least reliably be assigned. Of 2476 records, 38 were randomized trials and 17 were validation studies. No regulator or HTA body has yet been documented accepting or rejecting an AI-derived endpoint as such.

The measured gains are concentrated where ground truth is observable, which is to say in the clerical layer, and they are absent where the quantity of interest is a counterfactual. That division is stable across every workflow reviewed here and is the reason a scoping review of this field has to report its own error rate before it reports anything else. Part 2 asks what the regulatory and submission record says about the same division, and what reporting would make it auditable.

Disclosures: The author reports no conflicts of interest related to this work. No funding was received for this study. The author is affiliated with an academic institution and a health economics consultancy that conducts real-world evidence studies for pharmaceutical and life sciences clients; no client had any role in the conception, conduct, analysis, or reporting of this review, and no client reviewed the manuscript before submission. The large language model classifier described in the Methods is a measured object of study in this review rather than an assistive writing tool. Its prompts, inputs, and outputs are archived in the replication package. No AI tool is listed as an author.

Data Availability: All search strings, retrieved records, the classification dictionary, the blinded classifier prompts and outputs, the analysis scripts, the figure generation code, and the complete reference library are available in the public replication package (<https://github.com/obaser3791/ai-heor-scoping-review>). No individual patient data were used.

REFERENCES

1. US Food and Drug Administration. FDA Proposes framework to advance credibility of AI models used for drug and biological product submissions. Press announcement. January 6, 2025. October 6, 2026. <https://www.fda.gov/news-events/press-announcements/fda-proposes-framework-advance-credibility-ai-models-used-drug-and-biological-product-submissions>
2. Liu Q, Huang R, Hsieh J, et al. Landscape analysis of the application of artificial intelligence and machine learning in regulatory submissions for drug development from 2016 to 2021. *Clin Pharmacol Ther.* 2022;113(4):771-774. doi:10.1002/cpt.2668
3. Naik K, Goyal RK, Foschini L, et al. Current status and future directions: the application of artificial intelligence/machine learning for precision medicine. *Clin Pharmacol Ther.* 2024;115(4):673-686. doi:10.1002/cpt.3152
4. US Food and Drug Administration. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products. Draft guidance for industry. Docket FDA-2024-D-4689. Silver Spring, MD: FDA; January 2025. October 6, 2026. <https://www.fda.gov/media/184830/download>
5. US Food and Drug Administration and European Medicines Agency. *Guiding Principles of Good AI Practice in Drug*

- Development*. January 2026. October 6, 2026. <https://www.fda.gov/media/189581/download>
6. European Medicines Agency. EMA and FDA set common principles for AI in medicine development. News. January 2026. October 6, 2026. <https://www.ema.europa.eu/en/news/ema-fda-set-common-principles-ai-medicine-development-0>
 7. US Food and Drug Administration. Artificial intelligence-enabled medical devices. Device list current as of September 4, 2026. October 6, 2026. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices>
 8. International Society for Pharmacoeconomics and Outcomes Research. ISPOR top 10 HEOR trends. October 6, 2026. <https://www.ispor.org/heor-resources/about-heor/top-10-heor-trends>
 9. Fleurence RL, Wang X, Bian J, et al. A taxonomy of generative artificial intelligence in health economics and outcomes research: an ISPOR Working Group Report. *Value Health*. 2025;28(11):1601-1610. doi:10.1016/j.jval.2025.04.2167
 10. Fleurence RL, Dawoud D, Bian J, et al. ELEVATE-GenAI: reporting guidelines for the use of large language models in health economics and outcomes research: an ISPOR Working Group Report. *Value Health*. 2025;28(11):1611-1625. doi:10.1016/j.jval.2025.06.018
 11. Fleurence RL, Bian J, Wang X, et al. Generative artificial intelligence for health technology assessment: opportunities, challenges, and policy considerations: an ISPOR Working Group Report. *Value Health*. 2025;28(2):175-183. doi:10.1016/j.jval.2024.10.3846
 12. Use of artificial intelligence and machine learning in evidence synthesis for health technology assessment submissions: a review of NICE, CADTH, and IQWiG documents. Poster presented at: ISPOR 2026; Philadelphia, PA; May 17-20, 2026. October 6, 2026. <https://www.ispor.org/docs/default-source/cti-meeting-21463-documents/9774d614-3a98-4fba-bcf4-c1ab426db452.pdf>
 13. The use of artificial intelligence in scientific publishing. Editorial. *Lancet Infect Dis*. 2024. October 6, 2026. [https://www.thelancet.com/journals/laninf/article/PIIS1473-3099\(24\)00160-9/fulltext](https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(24)00160-9/fulltext)
 14. Lee W, Schwartz N, Bansal A, et al. A scoping review of the use of machine learning in health economics and outcomes research: part 1: data from wearable devices. *Value Health*. 2023;26(2):292-299. doi:10.1016/j.jval.2022.08.005
 15. Lee W, Schwartz N, Bansal A, et al. A scoping review of the use of machine learning in health economics and outcomes research: part 2: data from nonwearables. *Value Health*. 2022;25(12):2053-2061. doi:10.1016/j.jval.2022.07.011
 16. von GH, Helenius M, Hörhammer I, Moen H, Peltonen L. Scoping review on the economic aspects of machine learning applications in healthcare. *Int J Med Inform*. 2026;205:106103. doi:10.1016/j.ijmedinf.2025.106103
 17. Daghash H, Kernohan A, Brownson-Smith R, et al. Machine learning in health economic evaluations: protocol for a scoping review. *JMIR Res Protoc*. 2025;14:e77494. doi:10.2196/77494
 18. Wynants L, Van CB, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ*. 2020;m1328. doi:10.1136/bmj.m1328
 19. Andaur NCL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. 2021;n2281. doi:10.1136/bmj.n2281
 20. Baser O. Replication package for artificial intelligence in health economics and outcomes research: a systematic scoping review. GitHub repository. 2026. October 6, 2026. <https://github.com/obaser3791/ai-heor-scoping-review>
 21. Matsui K, Utsumi T, Aoki Y, Maruki T, Takeshima M, Takaesu Y. Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening: 3-layer strategy using GPT-3.5 and GPT-4 for systematic reviews. *J Med Internet Res*. 2024;26:e52758. doi:10.2196/52758
 22. Tran V, Gartlehner G, Yaacoub S, et al. Sensitivity and specificity of using GPT-3.5 Turbo models for title and abstract screening in systematic reviews and meta-analyses. *Ann Intern Med*. 2024;177(6):791-799. doi:10.7326/m23-3389
 23. Oami T, Okada Y, Nakada T. GPT-3.5 Turbo and GPT-4 Turbo in title and abstract screening for systematic reviews. *JMIR Med Inform*. 2025;13:e64682. doi:10.2196/64682
 24. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res Synth Methods*. 2024;15(4):616-626. doi:10.1002/jrsm.1715
 25. Nykvist B, Macura B, Xylia M, Olsson E. Testing the utility of GPT for title and abstract screening in environmental systematic evidence synthesis. *Environ Evid*. 2025;14(1):7. doi:10.1186/s13750-025-00360-x
 26. Flemmyng E, Noel-Storr A, Macura B, et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence. *Environ Evid*. 2025;14(1):20. doi:10.1186/s13750-025-00374-5
 27. Lee K, Paek H, Ofoegbu N, et al. A4SLR: an agentic artificial intelligence-assisted systematic literature review framework to augment evidence synthesis for health economics and outcomes research and health technology assessment. *Value Health*. 2025;28(11):1655-1664. doi:10.1016/j.jval.2025.08.002
 28. Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. *Res Synth Methods*. 2024;15(4):576-589. doi:10.1002/jrsm.1710
 29. Liu J, Lai H, Zhao W, et al. AI-driven evidence synthesis: data extraction of randomized controlled trials with large language models. *Int J Surg*. 2025;111(3):2722-2726. doi:10.1097/js9.0000000000002215
 30. Lai H, Liu J, Bai C, et al. Language models for data extraction and risk of bias assessment in complementary medicine. *npj Digit Med*. 2025;8(1):74. doi:10.1038/s41746-025-01457-w
 31. Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial intelligence to automate health economic modelling: a case study to evaluate the potential application of large language models. *Pharmacoeconomics Open*. 2024;8(2):191-203. doi:10.1007/s41669-024-00477-8
 32. Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial intelligence to automate network meta-analyses: four case studies to evaluate the potential application of large language models. *Pharmacoeconomics Open*. 2024;8(2):205-220. doi:10.1007/s41669-024-00476-9
 33. Sharp S, Lokuge K, Elvidge J, Hudson T, Dawoud D. Systematic literature review of the use of generative AI in health economic evaluation. 2025. doi:10.1101/2025.04.25.25326412
 34. Abogunrin S, Muir JM, Zerbin C, Sarri G. How much can we save by applying artificial intelligence in evidence synthesis? Results from a pragmatic review to quantify workload efficiencies and cost savings. *Front Pharmacol*. 2025;16:1454245. doi:10.3389/fphar.2025.1454245

35. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemp Clin Trials Commun.* 2019;16:100443. doi:10.1016/j.conctc.2019.100443
36. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology.* 2009;20(4):512-522. doi:10.1097/ede.0b013e3181a663cc
37. Guertin JR, Rahme E, LeLorier J. Performance of the high-dimensional propensity score in adjusting for unmeasured confounders. *Eur J Clin Pharmacol.* 2016;72(12):1497-1505. doi:10.1007/s00228-016-2118-x
38. Pang M, Schuster T, Filion KB, Eberg M, Platt RW. Targeted maximum likelihood estimation for pharmacoepidemiologic research. *Epidemiology.* 2016;27(4):570-577. doi:10.1097/ede.0000000000000487
39. Benasseur I, Talbot D, Durand M, et al. A comparison of confounder selection and adjustment methods for estimating causal effects using large healthcare databases. *Pharmacoepidemiol Drug Saf.* 2022;31(4):424-433. doi:10.1002/pds.5403
40. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat.* 2006;2(1). doi:10.2202/1557-4679.1043
41. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol.* 2007;6(1). doi:10.2202/1544-6115.1309
42. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc.* 2018;113(523):1228-1242. doi:10.1080/01621459.2017.1319839
43. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econometrics J.* 2018;21(1):C1-C68. doi:10.1111/ectj.12097
44. Bonander C, Svensson M. Using causal forests to assess heterogeneity in cost-effectiveness analysis. *Health Econ.* 2021;30(8):1818-1832. doi:10.1002/hec.4263
45. Zhang Y, Kreif N, GC VS, Manca A. Machine learning methods to estimate individualized treatment effects for use in health technology assessment. *Med Decis Making.* 2024;44(7):756-769. doi:10.1177/0272989x241263356
46. Glynn D, Giardina J, Hatamyar J, Pandya A, Soares M, Kreif N. Integrating decision modeling and machine learning to inform treatment stratification. *Health Econ.* 2024;33(8):1772-1792. doi:10.1002/hec.4834
47. Möller M, Wild E, Tan W, Schreyögg J. Estimating heterogeneous treatment effects with real-world health data: a scoping review of machine learning methods. *Value Health.* 2026;29(5):905-913. doi:10.1016/j.jval.2026.01.013
48. Manke-Reimers F, Brugger V, Bärnighausen T, Kohler S. When, why and how are estimated effects transported between populations? A scoping review of studies applying transportability methods. *Eur J Epidemiol.* 2025;40(3):255-273. doi:10.1007/s10654-025-01217-w
49. Turner AJ, Sammon C, Latimer N, et al. Transporting comparative effectiveness evidence between countries: considerations for health technology assessments. *Pharmacoeconomics.* 2023;42(2):165-176. doi:10.1007/s40273-023-01323-1
50. Weise A, Büchter RB, Pieper D, Mathes T. Assessing transferability in systematic reviews of health economic evaluations, a review of methodological guidance. *BMC Med Res Methodol.* 2022;22(1):52. doi:10.1186/s12874-022-01536-6
51. Wang K, Rosman LA, Lu H. Machine learning versus logistic regression for propensity score estimation: a benchmark trial emulation against the PARADIGM-HF Randomized Trial. 2025. doi:10.1101/2025.06.16.25329708
52. US Food and Drug Administration. *Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products.* Guidance for industry. Docket FDA-2020-D-2307. July 2024. October 6, 2026. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory>
53. Gruber S, Phillips RV, Lee H, Concato J, van der Laan MJ. Evaluating and improving real-world evidence with targeted learning. *BMC Med Res Methodol.* 2023;23(1):178. doi:10.1186/s12874-023-01998-2
54. Padula WV, Kreif N, Vanness DJ, et al. Machine learning methods in health economics and outcomes research: the PALISADE Checklist: a good practices report of an ISPOR Task Force. *Value Health.* 2022;25(7):1063-1080. doi:10.1016/j.jval.2022.03.022
55. Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc.* 2021;28(12):2716-2727. doi:10.1093/jamia/ocab170
56. Kharrazi H, Anzaldi LJ, Hernandez L, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. *J Am Geriatr Soc.* 2018;66(8):1499-1507. doi:10.1111/jgs.15411
57. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. *npj Digit Med.* 2024;7(1):6. doi:10.1038/s41746-023-00970-0
58. Xu W, Gu B, Lotter WE, Kehl KL. Extraction and imputation of Eastern Cooperative Oncology Group performance status from unstructured oncology notes using language models. *JCO Clin Cancer Inform.* 2024:e2300269. doi:10.1200/cci.23.00269
59. Maghsoudi A, Sada YH, Nowakowski S, et al. A multi-institutional natural language processing pipeline to extract performance status from electronic health records. *Cancer Control.* 2024;31:10732748241279518. doi:10.1177/10732748241279518
60. Lin E, Zwolinski R, Wu JT, et al. Machine learning-based natural language processing to extract PD-L1 expression levels from clinical notes. *Health Informatics J.* 2023;29(3):14604582231198021. doi:10.1177/14604582231198021
61. Kenney RC, Chen X, Shintani K, et al. Validation of non-small cell lung cancer clinical insights using a generalized oncology natural language processing model. *JCO Clin Cancer Inform.* 2024:e2300099. doi:10.1200/cci.23.00099
62. Abedian S, Sholle ET, Adekanattu PM, et al. Automated extraction of tumor staging and diagnosis information from surgical pathology reports. *JCO Clin Cancer Inform.* 2021;1054-1061. doi:10.1200/cci.21.00065
63. Varma G, Yenukoti RK, Kumar MP, et al. A deep learning-enabled workflow to estimate real-world progression-free survival in patients with metastatic breast cancer: study using deidentified electronic health records. *JMIR Cancer.* 2025;11:e64697. doi:10.2196/64697
64. Estevez M, Singh N, Dyson L, et al. Ensuring reliability of curated electronic health record-derived data: the validation of accuracy for large language model-/machine learning-extracted information and data (VALID) framework. *JCO Clin Cancer Inform.* 2026;10(1):e2500215. doi:10.1200/cci-25-00215
65. Flatiron Health. Flatiron Health publishes first peer-reviewed validation framework for AI-extracted real-world oncology data.

- Press release. October 6, 2026. <https://resources.flatiron.com/press/flatiron-health-publishes-first-peer-reviewed-validation-framework-for-ai-extracted-real-world-oncology-data-in-journal-of-clinical-oncology>
66. Ma X, Bellomo L, Magee K, et al. Characterization of a real-world response variable and comparison with RECIST-based response rates from clinical trials in advanced NSCLC. *Adv Ther*. 2021;38(4):1843-1859. doi:10.1007/s12325-021-01659-0
 67. Development and validation of a machine learning approach to scale real-world response in oncology. Poster presented at: ISPOR. October 6, 2026. <https://www.ispor.org/docs/default-source/cti-meeting-21021-documents/6f2aef20-5193-4ad1-b198-1156a7ea80f7.pdf>
 68. Flatiron Health. *Flatiron Health Research on AI-Driven Cancer Progression Extraction Presented at AACR Special Conference in Cancer Research: Artificial Intelligence*. Press release. October 6, 2026. <https://resources.flatiron.com/press/flatiron-health-research-on-ai-driven-cancer-progression-extraction-presented-at-aacr-special-conference-in-cancer-research-artificial-intelligence>
 69. Riskin DJ, Monda KL, Gagne JJ, et al. Implementing accuracy, completeness, and traceability for data reliability. *JAMA Netw Open*. 2025;8(3):e250128. doi:10.1001/jamanetworkopen.2025.0128
 70. Truveta. *Truveta Studio and Data Quality*. Presentation at: ISPOR Europe 2024. October 6, 2026. <https://www.ispor.org/docs/default-source/euro2024/ispor-eu-2024---truveta-final.pdf>
 71. Pfizer. *US FDA Approves IBRANCE (Palbociclib) for the Treatment of Men With HR+, HER2- Metastatic Breast Cancer*. Press release. October 6, 2026. <https://www.pfizer.com/news/press-release/press-release-detail/u-s-fda-approves-ibrance-palbociclib-for-the-treatment-of-men-with-hr-her2-metastatic-breast-cancer>
 72. Degeling K, IJzerman MJ, Lavieri MS, Strong M, Koffijberg H. Introduction to metamodeling for reducing computational burden of advanced analyses with health economic models: a structured overview of metamodeling methods in a 6-step application process. *Med Decis Making*. 2020;40(3):348-363. doi:10.1177/0272989x20912233
 73. Fleurence R, Bian J, Wang X, et al. *Generative AI for health technology assessment: opportunities, challenges, and policy considerations*. arXiv. 2024. doi:10.48550/arxiv.2407.11054
 74. Hasan B, Saadi S, Rajjoub NS, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. *BMJ Evid Based Med*. 2024;29(6):394-398. doi:10.1136/bmjebm-2023-112597
 75. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models “the answer” to bringing us closer to systematic review automation? *Syst Rev*. 2023;12(1):72. doi:10.1186/s13643-023-02243-z
 76. Otten TM, Grimm SE, Ramaekers B, Joore MA. Comprehensive review of methods to assess uncertainty in health economic evaluations. *Pharmacoeconomics*. 2023;41(6):619-632. doi:10.1007/s40273-023-01242-1
 77. Bojke L, Claxton K, Sculpher M, Palmer S. Characterizing structural uncertainty in decision analytic models: a review and application of methods. *Value Health*. 2009;12(5):739-749. doi:10.1111/j.1524-4733.2008.00502.x
 78. Fenwick E, Steuten L, Knies S, et al. Value of information analysis for research decisions: an introduction: report 1 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. *Value Health*. 2020;23(2):139-150. doi:10.1016/j.jval.2020.01.001
 79. Rothery C, Strong M, Koffijberg H, et al. Value of information analytical methods: report 2 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. *Value Health*. 2020;23(3):277-286. doi:10.1016/j.jval.2020.01.004
 80. Sculpher M, Claxton K. Establishing the cost-effectiveness of new pharmaceuticals under conditions of uncertainty: when is there sufficient evidence? *Value Health*. 2005;8(4):433-446. doi:10.1111/j.1524-4733.2005.00033.x
 81. Fenwick E, Claxton K, Sculpher M. The value of implementation and the value of information: combined and uneven development. *Med Decis Making*. 2007;28(1):21-32. doi:10.1177/0272989x07308751
 82. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) Statement: updated reporting guidance for health economic evaluations. *Value Health*. 2022;25(1):3-9. doi:10.1016/j.jval.2021.11.1351
 83. D’Amico S, Dall’Olio D, Sala C, et al. Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. *JCO Clin Cancer Inform*. 2023:e2300021. doi:10.1200/cci.23.00021
 84. Elvatun S, Knoors D, Brant S, Jonasson C, Nygård JF. Synthetic data as external control arms in scarce single-arm clinical trials. *PLOS Digit Health*. 2025;4(1):e0000581. doi:10.1371/journal.pdig.0000581
 85. Wang D, Florian H, Lynch S, et al. Using AI-generated digital twins to boost clinical trial efficiency in Alzheimer’s disease. *Ac&D Transl Res Clin Interv*. 2025;11(4):e70181. doi:10.1002/trc2.70181
 86. Akiya I, Ishihara T, Yamamoto K. Comparison of synthetic data generation techniques for control group survival data in oncology clinical trials: simulation study. *JMIR Med Inform*. 2024;12:e55118. doi:10.2196/55118
 87. Kaabachi B, Despraz J, Meurers T, et al. A scoping review of privacy and utility metrics in medical synthetic data. *npj Digit Med*. 2025;8(1):60. doi:10.1038/s41746-024-01359-3
 88. Kopanitsa G. Validation is not enough: longitudinal evidence of post-deployment fragility in clinical AI systems. *PLOS Digit Health*. 2026;5(7):e0001534. doi:10.1371/journal.pdig.0001534
 89. Artificial intelligence tools in health technology assessment: a scoping review. Conference proceeding OP07. October 6, 2026. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11718724/>